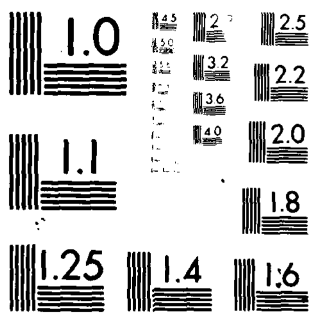
AD-A092 664   MASSACHUSETTS INST OF TECH CAMBRIDGE MAN-MACHINE SYS--ETC F/G 5/8
SUBJECTIVE MEASUREMENT OF MENTAL WORKLOAD.(U)
MAY 80 N MORAY   N00014-77-C-0256

UNCLASSIFIED   NL

1 OF 1
AD
A092664

END
DATE
FILMED
1 81
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

LEVEL II 12 b.s

AD A092664

SUBJECTIVE MEASUREMENT OF MENTAL WORKLOAD

NEVILLE MORAY

May 1980

Contract N00014-77-C-0256
Engineering Psychology Programs
Office of Naval Research
Arlington, VA 22217

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO. AD-A092664

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
SUBJECTIVE MEASUREMENT OF MENTAL WORKLOAD.

5. TYPE OF REPORT & PERIOD COVERED
Technical Report.
1 July 1979 - 30 June 1980.

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Neville Moray

8. CONTRACT OR GRANT NUMBER(s)
N00014-77-C-0256

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR196-152

11. CONTROLLING OFFICE NAME AND ADDRESS
Engineering Psychology Programs
Office of Naval Research (Code 455)
Arlington, VA 22217

12. REPORT DATE
May 1980

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
NONE

Accession For
NTIS GRA&I
A

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
workload
mental workload
stress
human performance
subjective judgement
engineering psychology
man-machine systems

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Although there is widespread agreement that an important component of mental workload is the subjective judgement of how difficult the task seems, and how loaded the human operator feels, there have been rather few attempts to measure subjective load directly. This paper reviews those attempts and discusses what variables in the task seem to be responsible for the ratings given by the operator. Some suggestions for future research are given.

Introduction

Despite the fact that the term <u>mental workload</u> is not indexed in Psychological Abstracts, it has come to have increasing currency in applied psychology in recent years because it reflects a real feature of man-machine interaction. At least three symposia have been held on the topic in the last three years, and an extensive review was published by Wierwille and Williges (1978). Attempts have been made to measure mental workload in a variety of ways which have been summarized, for example, as in Moray (1979) and a recent issue of the <u>Human Factors</u> journal (1979). Generally speaking, there are four categories which have become accepted to describe workload: physiological measures, performance measures, task variable measures, and subjective measures. This paper is concerned only with the last of these. It deals only with research in which the subject or operator has been asked directly to estimate how difficult or how loading the task feels to him.

The matter is of considerable practical importance. As man-machine systems become more and more complex and automatic control becomes more sophisticated, tasks require less and less physical exertion by the operator, but are still experienced as difficult and exhausting. From the point of view of setting appropriate wage rates, of safety, and of efficiency it would be highly desirable to know what aspects of a task make it seem difficult, even under conditions where performance is perfect. For it is clear that there will be times when perfect performance is achieved only by the operator working harder and harder until a point of breakdown is approached. Schmidt (1978) has suggested a compliance analogy: there may be no deformation in a stiff structure until very near the point at which it fails. Clearly it would be of great value to know how near the breaking point the human operator might be at any time.

We will, therefore, examine only published empirical measures of subjective mental load (SML) since 1968. References to earlier work can be found, for example, in Roscoe, Ellis and Chiles (1978).

There is no agreed definition of mental workload and no agreement as to how to measure it. Many writers (e.g. LePlat, 1978) maintain that it should be tied to task variables, to personality variables, to physical or physiological variables, and also to social variables such as social pressure and expectations. An objectively easy task may feel hard although it is performed perfectly due to fatigue, to payoffs associated with various outcomes, and to the motivational state of the operator. On the other hand appropriate instructions and a suitable balance between speed and accuracy may make an objectively difficult task seem easy. It is necessary to make the distinction between the mental load called for by the physical parameters of the task, which we will here call imposed mental load (IML), and subjective mental load (SML).

Data are astonishingly sparse. Despite lip service paid to its importance, such that some writers maintain the SML is the only real meaning of mental workload (see e.g. papers in Moray, (1979)) there seem not to have been more than a handful of papers published in the last ten years, and of those the vast majority are concerned with manual control tasks. Almost none have been published in connection with monitoring tasks.

It will be assumed that terms such as SML, "perceived difficulty", "perceived effort", etc. are sufficiently closely related to be taken to refer to a single phenomenon. It should, however, be emphasized that we do not know this for certain. It is common, for example, to find SML being said to be due to the amount of effort expended, neither term being defined.

Measurements of SML may conveniently be divided into four groups: measures related to physical and physiological task parameters; measures related to cognitive tasks; measures related to manual control tasks, generally tracking tasks or aircraft control; and measures related to "time stress". Each category will be examined in turn.

Measures Related to Physical and Physiological Parameters

An extensive program of research in this area has been undertaken by Borg and his colleagues at the Stockholm Institute of Applied Psychology. Both empirical and theoretical studies into scaling have been undertaken (Borg, 1971; 1978a; 1978b; Borg, Bratfisch, and Dornic, 1971; Hallsten and Borg, 1975). With regard to scaling, their results suggest that the kind of scale used is not very critical, the choice of category scales, magnitude estimation, and the presence or absence of verbal labels making little difference.

Borg (1978a, 1978b) examined the perceived difficulty of physical tasks. These included lifting weights, wheeling a wheelbarrel, and cycling on a bicycle ergometer. His overall conclusion is that the relation between the intensity of physical work and the perceived difficulty of the work is a power function:

$$D = a + c(S - b)^{n}$$

where 'a' is a 'perceptual noise constant' (actually just the point at which the curve crosses the ordinate at S = 0 and not to be confused with the 'perceptual noise' mentioned later in connection with Optimal Control Theory), 'c' is a scaling constant, 'b' a constant determined by the kind of work, 'S' the physical intensity of the work, and 'n' the exponent. For example, for heavy physical work on the ergometer Borg gives:

$$D = 4 + 0.001(S - 0)^{1.6}$$

Most of the exponents are in the range 1.2 - 1.7 with S measured in km/hr , although in the case of walking on a treadmill the exponent is as high as 3.0.

It is not known for certain what it is that a human senses when making judgments of difficulty. In physical tasks it seems as though heart rate could be the variable sensed by the participant and used to estimate task difficulty, since there are very strong correlations between heart rate and ratings of perceived difficulty. Some workers (see Johannsen et al., 1979) have suggested that all SML is secondary to physiological events such as heart rate changes, muscle tension, etc. On the other hand, Borg's work suggests that participants can distinguish the source of the difficulty with some precision, and in particular can distinguish between cardiovascular sensations and peripheral muscular sensations. In one experiment heart rate and perceived difficulty were related to weights lifted by the arm. Heart rate, overall, perceived difficulty, and perceived difficulty of arm movements were highly correlated, but the perceived

difficulty of leg movements was independent. When cycling on the ergometer on the other hand 'perceived exertion' of the arms had a much slower rate of growth with physical intensity measured in km/hr than did overall perceived exertion or perceived leg exertion.

This raises the interesting possibility that a human operator may in a quite general way be able to distinguish the specific sources of difficulty, and to rank order them, in a multidimensional task, rather than merely giving an overall rating of task difficulty. This does not seem to have been attempted. Such an ability could be a valuable source of information in the design of operator stations in complex man-machine systems.

Borg switches between the terms 'perceived difficulty' and 'perceived exertion' without considering the problem of definition. But there is a real question as to whether the difficulty of a physically demanding task is closely related to the difficulty of a mentally demanding task. To assume so seems to mean that in some way the participant examines the task and then reports on his feelings of adequacy rather than on the feelings induced directly by the task. Alternatively one might assume that there were two variables which are described by the word 'difficulty' which are highly correlated. No attempt seems to have been made to investigate this point.

Two studies fall somewhere between physiological and cognitive tasks. These are investigations of fatigue by Yoshitake (1971) and Stave (1977). The former investigated the relation between feelings of fatigue and symptoms of fatigue in bank personnel. He asked participants to rate their feelings of fatigue on a 9 point scale from 'feeling fit, rested' to 'feeling extremely tired, exhausted'. He found a high correlation (r = + 0.8, or better) between the occurrence of such symptoms as an unwillingness to think, lack of confidence, and failure of concentration and the ratings of subjective fatigue. There was also a strong correlation, albeit somewhat smaller, with physical symptoms such as yawning, stiffness of the neck, etc. Stave (1977) examined helicopter pilots flying simulator sorties of several hours duration. He found that the occurrance of mental 'blocks' correlated + 0.87 with self ratings of fatigue, and related the blocks to the findings of Bills (1931) on the effects of prolonged work. Stave's reports are very reminiscent of the 'microsleeps' reported by Oswald (1962). The subjective ratings stayed almost constant until near the end of the mission. When the pilots knew that the end was near, fatigue began to rise rapidly, suggesting that factors such as expectancy and motivation may play a major part in the onset of subjective fatigue.

In the context of physiological work it is interesting that there seems to have been little attempt to make use of the traditional assertion that there is an 'inverted U-shaped curve' relating physiological arousal to the efficiency of performance. The only direct study seems to be one by Verplank (1977) who looked for the optimum level of imposed workload to maximize performance. He failed to find any such optimum, which could have been related to the U-shaped curve, and found instead a linear decrease in performance with increasing workload. Tulga (1978) by contrast did find an inverted U-shaped relation between performance and subjective workload in a supervisory task, but there is no reason to think that it was due to arousal. Rather he related it to what was essentially a kind of speed accuracy tradeoff. As load increased so did performance until the operator's information processing ability was exceeded. After that the operator became less accurate thus showing reduced performance. He also felt there to be less subjective workload despite the rising imposed workload, presumably due to his adopting a less strict criterion of what was satisfactory performance.

Measurement of SML in Cognitive Tasks

Here again, the Stockholm group played a major role. They gave their participants a set of intellectual tasks of varying difficulty, and measured the frequency of solution and speed of solution. Scores were transformed so that the task with a probability of solution of 0.5, or the nearest to the mean time for solution, was given a score of zero, and the scores for the other tasks were converted to standard scores. The performance tasks were therefore scaled to zero mean and unit standard deviation, and the ratings of perceived difficulty correlated with that scale. Either a 9-point scale or the method of magnitude estimation were used.

Borg, Bratfisch, and Dornic (1971) examined the subjective difficulty of a visual search task. Pairs of consonants such as DH, LV, etc. were displayed in the cells of matrices of order 5x5 to 11x11. The 8x8 matrix was used as a standard. Magnitude estimation of difficulty was used the standard being assigned the number 10. The objective score was the time to find the designated target pair. Perceived task difficulty was found to relate linearly to the logarithm of the number of stimuli and to search time. It is not clear what aspect of the task the participants were rating, but they were asked for a global estimate, and to avoid using the length of time it took them to reach a solution or other such intervening variables. An interesting observation was that those participants who scored highest on a Raven intelligence test rated any given problem higher in difficulty than those with lower Raven scores although there was no difference in the time taken to solution.

In another study the same workers (Bratfisch et al., 1972a) examined the subjective difficulty of several tasks from a Swedish standardized intelligence test. The first task was to find the rule governing a number series and to supply the next two members of the series. The second was to look at a diagram showing a series of linked levers, and to say in which direction an indicated lever would move when another specified lever was moved. The third task was to choose a synonym for a word from a set of alternatives provided. In this case the category scale was used. A Spearman rank order correlation of +0.9 was obtained between the objective and subjective task difficulty. Rankings at the extremes (1 and 9) appeared to be about $\pm$ 2.0 $\underline{z}$ on the normalized scale of objective difficulty. In the number series there were a total of 25 occasions on which the problem was not solved; and in each case it was rated 9, 'very, very difficult'. On being shown the answer, or the method of solution, the participants lowered their ratings to 7 or 8. In other words, a problem one cannot do is by definition 'very, very difficult', but when shown how to do it, it neither becomes 'easy' nor stays 'very, very difficult'. The regression equation of perceived difficulty on objective difficulty was:

$$D = 5.3 + 1.4(\underline{z}) \qquad 0 \leq D \leq 9$$

and changed only slightly for the different kinds of problems.

Again, Bratfisch et al. (1972b) found a strong relation between magnitude estimation of perceived difficulty and the objective difficulty of Raven's matrices. Two methods were used. In one the item with a probability of solution of 0.5 was assigned the number 10 as a standard, and in the other the item thought by the participant to be the most difficult was assigned 100. The mean estimates of the two procedures correlated +0.98 although the patterns of standrd deviations of the estimates of difficulty were different. The relation between perceived and objective difficulty was slightly curvilinear, and although no statistical tests appear to have been used the authors suggest that the relationship is best expressed as:

$$D = 10.54(1.54^{\underline{z}})$$

where $\underline{z}$ is the normalized objective score. Once again there was a difference in rating between the worst and the best participants. The best performers showed a steeper increase in perceived difficulty with objective difficulty than did the worse performers. This seems at first glance counter-intuitive and taken together with the study cited earlier suggests that the intellecutally more able tend to reflect on the quality of the task rather than plunging in and attempting to solve it. Although the evidence is very slight, being confined to these two experiments, it would be worth more research into the degree with which high intelligence is associated with a tendency to see difficulties. If a task is rated more difficult than it should be by someone who is able to perform it, it is possible that his judgment could have an adverse effect on his motivation.

The Stockholm group has carried out a considerable number of these investigations, which are summarized by Borg (1978b). In all cases they were able to achieve high correlations between the subjective and objective measures of difficulty. There is no doubt, therefore, that the notion of 'subjective difficulty' is one which people can use consistently. But what is it that they are using as the basis of a judgment which can cover the range from pushing a wheelbarrow to solving a difficult problem in a Raven matrix?

Borg is not particularly helpful on this point.:

> "If some difficult mental work is compared with hard muscular work, there is in both cases a stimulus situation that most often may be fairly well defined. There is a subject who works on a task and reacts with a complex response syndrome depending upon stimulus - response variables as

> influenced by the subjects' abilities, experience, pesonality traits, etc...That mental tasks are different as regards difficulty can mean either that they are differently difficult to solve, or that they are experienced as differently difficult, or maybe that there are differences in some relevant physiological aspects...If we make a comparison with physical work, the concepts of force, work and power are of special interest. Force is physical intensity equal to kp, work is the amount of force exerted for some distance, that is kpm, and finally, power is work per unit time, kpm/min (watt). Similar concepts might be used in mental work. We may momentarily exert a certain "force" to solve a problem, we may use this force for a certain time and thus do a certain amount of mental work, and we may also study the mental power, that is, the mental efficiency."

The problem here is that there is no mental force which is operationally defined and internally dimensionally consistent. In physical units force is defined as (mass.acceleraion), but to speak of mass or acceleration in mental terms is simply meaningless. Moreover, we cannot arbitrarily substitute time for distance in calculating work, and it is simply incorrect to equate power consumption with efficiency. This tendency to carry over units from one domain to another without adequate operational definitions and internal consistency is rife in psychology and causes nothing but confusion and pseudo-explanations. (The same objection can be made to the general use of "effort", which, although undefined and unmeasured, has come to have a great popularity in the last few years as a so-called explanation.) Less misleading, but equally difficult to use quantitatively, is Borg's suggestion that the judgment of difficulty is related to:

> "....a confrontation of the present task with the content of one's long term memory storage including both general experience and memories of similar tasks.....background factors such as personality traits, habits, likes and dislikes, aspirations and expectation levels....one's emotional state, general fatigue, ...motivation....the importance one ascribes to the task ....anticipated success or failure."

This is no doubt true, but not very helpful. If it is true that all the above variables really condense onto a single dimension of "perceived difficulty" it will be a remarkable s roke of luck, considering the efforts which have gone into testing them as separate dimensions of personality in the past.

Two further points should be noted. One is that we have made an implicit assumption that what is perceived as difficult is perceived as producing SML: no experiment has actually related the two judgments. The second is that unlike the tasks in the next two sections these tasks are all single trial tasks. The participant is not under the pressure associated with a continuous or arbitrary stream of signals whose members may arrive before he has finished dealing with an earlier signal. He was not, in the above experiments, under any kind of time stress, nor did his response in any way determine what would happen next.

Measurements of SML in Manual Control Tasks

The basis of measurement of mental load in these tasks are the Cooper and Cooper- Harper scales (Cooper, 1957; Cooper and Harper, 1969). These scales were developed to measure the handling characteristics of aircraft by using the subjective reports of test pilots. They are well established and validated, and have been in extensive use for 20 years. Their mathematical properties have been examined (McDonnell, 1969) and they can be regarded as useful and reliable instruments for determining the "flyability" of aircraft. There is a very large number of reports in which they have been so used. In the present context, however, it is necessary to concentrate on the very much smaller data base in which the ratings throw light not on the aircraft but on how the pilot makes his judgment as to what rating on the C or CH scale to give. As with Borg's work we must make an assumption. In this case it is that if a pilot states that an aircraft is difficult (or impossible) to fly, this is equivalent to his saying that the task of flying it imposes on him a very heavy load (or an impossible load, an overload) and so on down the scale. With one exception (Wewerinke, 1974) no one has directly compared judgments of SML and C or CH ratings. Considering the success of the CH scale it is rather surprising that no attempt has been made to generalize it to other applications (but see Moray, 1979b; Sheridan, 1979). McDonnell (1969) shows in detail how this could be done.

Use of the CH scale may conveniently be thought of as falling under three subheadings. It has been used in connection with models derived from classical control theory by McRuer and his coworkers; with models derived from Optimal Control Theory (OCT) by Baron, Kleinman and Levison; and in the direct application for which it was designed, namely to assess flight characteristics. In reviewing this work, the real point of interest is to establish what characteristics of manual control tasks give rise to the scaled judgments. What is it about the task that drives the CH rating up towards the heavily loaded end of the scale?

It is convenient to begin with Wewerinke's (1974) paper already mentioned, since from it we may conclude that the identification of CH ratings with mental load is justified. Wewerinke took three scales from McDonnell's paper (McDonnell, 1968). One was the original Cooper scale, one was a parallel scale but with the wording adjusted specifically to ask how hard the task seemed to the operator, and the third was a ten point scale with no verbal labels, on which the operator was to mark the point corresponding to the degree of difficulty of the task. (See Figure 1). The operators performed a compensatory tracking task, with 6 different controlled elements, k, k/s, k/s , k/s(s-0.5), k/(s-3) and k/s(s-1). Objective difficulty was computed using the observation noise ratio as an a priori measure

of the attentional workload imposed by the task (see Levison, 1979, for a very clear introduction to these concepts). The subjective scores were normalized to a mean of zero and a standard deviation of 1.0, and averaged across operators. Several other OCT parameters such as transport delay, motor noise, etc. were also estimated for each operator.

The correlation between the verbal and the numerical 10-point scale was almost perfect (r = + 0.99) and thereafter only the numerical scale was used. The correlation between subjective difficulty rating and the objective workload defined in terms of OCT observation noise was also extremely high, (r = + 0.99). We may reasonably conclude that CH and C ratings do indeed relate closely to workload. Because OCT relates computed workload to observation noise Wewerinke proposed that the origin of subjective load is probably in the ratio of the error variance to the observation noise. (See Figure 2.)

Figure 2 also shows another well established fact about subjective load. There is complete agreement that the higher the order of control of the controlled element, and the more lead that must be generated by the human operator, the higher also the experienced load. Ashkenas (1966) reviewed many studies to that date and concluded that there was a linear relation between decrements in the C rating and the lead time constant required of the pilot, (Figure 3). McRuer and Weir (1969), Hess (1977), Hess and Teichgraber (1979), Jex and Clement (1979), and Arnold, Johnson, and Dillow (1973) all support this conclusion. Lead time constants of about 5 seconds, and second order control are the upper limits of human performance except in very special circumstances and with exceptional pilots.

Another aspect of manual control which affects subjective load is heterogeneity of control. If two processes must be simultaneously controlled but the ocntrol law required is identical, then there is little effect either on performance or on subjective load of having two tasks rather than one. But if a different control law is required for each, then adding the second increases SML (Wickens and Tsang, 1979).

A second universally accepted conclusion is that system instability increases SML. This is apparent in Figure 2 above. In Wewerinke's graph points (1), (2), and (4) require increasing amounts of lead, while points (3), (5) and (6) are for plants with increasing instability. McRuer and Weir (1969) found that the C rating depended on the gain and phase margins in the crossover region of the system Bode plot. And Ashkenas (1966) found that a Cooper rating of 2 could be driven up to 6 or more by reducing the damping coefficient in a second-order system below 0.8 (see figure 3). He also found an equation relating Dutch Roll characteristics of an aircraft to C ratings which

suggested that high ratings were produced by the experience of rapid acceleration.

The laboratory task which has been most extensively used to investigate the relation between SML and instability is the "critical task" designed by Jex (Jex and Clement (1979), Jex (1979)). The controlled element has a transfer function of the form which is an inherently unstable divergent element. The critical task changes the value of during a run, thus increasing the rate at which the divergence occurs, and the highest value of at which the operator can maintain control is taken as his objective score on the critical task.

A most convincing summary of the evidence on this point can be found in Jex and Clement (1979). From their data it is possible to estimate the relation between C ratings and , using a figure from McDonnell (1968). Because of the very consistent findings on the critical task, this provides a useful performance test which correlates perfectly with SML, the relation being approximately:

$$\text{Cooper rating} = 9 - 1.5\,\lambda \qquad (0 \leq \lambda \leq 6,\ \lambda \text{ in Hz}).$$

In addition Hess (1977) found a correlation of + 0.73 between λ/λ_c and a numerical scale similar to that used by Wewerinke, where λ_c the value of λ at which control is lost. McRuer and Weir (1969) and Hess and Teichgraber (1979) give similar results.

Several authors relate SML to the magnitude of the error signals observed by the operator. Wewerinke suggests it explicitly, and all OCT work implies it. Arnold, Johnson and Dillow (1973) were able to predict pilot ratings from error power and lead time constants, and found that the workload rating given by pilots was:

$$\text{Rating} = \text{PERF} + 0.43T_L + 1.0$$

$$\text{PERF} = 5.8\,\sigma_\theta + 0.43\,\sigma_\delta$$

where σ_θ and σ_δ are the rms errors of the two controlled variables and T_L is the lead time constant. Hess and Teichgraber (1974) suggested that SML can be predicted from the ratio of the error to forcing function power. They also showed that if analogue errors are coarsely quantised before being displayed SML increases and λ_c falls. Moreover, they found that the effect of coarse quantising was a separate factor which varied independently of lead generation. (In their paper direct estimates of SML were not obtained, but these conslusions stand because of the very strong relation (noted above) between SML and

critical task parameters.)

The use of OCT emphasises the observation signal-to-noise ratio, and hence the efficiency with which the observer can estimate the value of the system state variables. Hess (1977) examined the relation between CH ratings and J, the OCT index of performance, and found that:

$$CH = 1 \quad , \quad J < 1$$

$$Ch = J \quad , \quad 1 < J < 8$$

$$CH = 9 \quad , \quad 8 < J < \infty$$

$$CH = 10 \quad , \quad J = \infty$$

and went on to ask what aspect of J it was that leads to the particular value of CH. He concluded that the crucial variable is the fraction of attention allocated to the task, and suggested that the CH scale should be regarded as a direct measure of mental workload, and that it was unnecessary to ask therefore how CH and SML are related. (See Figures 4 and 5.) The general theory of attentional demand and allocation in OCT terms will be found in Levison (1979).

All the above studies have used either laboratory tracking tasks or pilots flying aircraft or simulators. Hess (1977) suggested that the Cooper-Harper scale could be applied when suitably worded to any control task, providing certain conditions were fulfilled. These are that the operator can observe directly all the relevant variables, and that in computing the index of performance and estimation model parameters the weighting coefficients on the index of performance are the reciprocals of the permissible error on the variables concerned.

There seems to be only one published report of the use of the CH scale with a vehicle other than aircraft. Rule and Fenton (1972) used the CH scale to measure the efficacy of a tactile aid to automobile driving in a task which required the driver to maintain station at a fixed distance from another vehicle. They found that when the system had a short lag time constant the headway error was reduced, but the CH rating increased by 1 unit, suggesting that the driver was having to work harder. This rise in CH rating could be offset by providing better information about the distance between the vehicles or by providing feed-forward in the control system. The inclusion of an "automatic pilot" also reduced the CH rating.

To end this section on control and SML we may note that McRuer and Krendel (1974) quoting Ashkenas (1972) give equations allowing separate ratings in a multi-loop task to be combined

into an overall rating, the inverse problem of that mentioned above where we saw that from the work of Borg we may be able to ask operators to fractionate their SML. The rule for combining ratings is:

$$R = A + \frac{1}{B^{(m-1)}} \cdot \prod^{m} (R_i - A)$$

where m is the number of subtasks, R_i the rating on the ith subtask, A=10 set by the limit of the CH scale, and B is an empirical constant, B = -8.3.

Measurements of SML as "Time Stress"

In so-called "real life" tasks other than manual control tasks operators often receive signals which must be processed before they have finished processing an earlier signal. This is an example of "time stress". Almost all reports on air traffic control, process control, nuclear power plant control, etc., refer to the problem of time stress and subjective mental workload, but remarkably few attempts have been made to measure the relation between them. Senders (personal communication) has gone so far as to assert that unless there is time stress in a task there is by definition no SML, and the popularity of time-line analysis as an engineering design tool implicitly supports that point (Parks, 1979).

The obvious conceptual framework for "time stress" is queueing theory, which predicts SML from the probability that the "server" (the human operator) will be busy when the "customer" (a signal or message) arrives. Recently Chu (1979) related SML to queueing theory parameters and obtained a correlation of +0.97 between the probability of server occupancy and the subjective effort rating. With each variable normalized to a range from 0 to 1 the data fell almost exactly on a straight line through the origin with a slope of 1.0. This strongly supports the queueing theory approach to SML.

Earlier studies include one by Philipps, Reiche and Kirchner (1971), who studied Air Traffic Controllers for periods of several hours. In addition to making a complete audio and video recording of each session, a record was kept of "stress of time" and "difficulty of control task". While the ATC performed his task another experienced controller observed him, and punched in ratings on the above variables on two ten-key keyboards, also making verbal comments. Thus the experiment was, strictly speaking, a judgment by an observer of "what I think I would feel if I were doing the task he is doing". However the record was played back to the active controller and a correlation of $r = +0.51$, p 0.01 was obtained between the ratings by the controllers. The objective load was computed as "bits of information per four minutes". Although this is not completely clarified in the paper, it appears to be an estimate of the information content of the dialogue exchanged between the controller and the aircraft. It is described only as "information content of dialogue groups in radio communication and of the processed control strips".

The main findings were:

1. Stress of time and rated difficulty gave a Spearman correlation of r = +0.51,(p 0.01).

2. There was no significant correlation between the information content of an episode (four minutes) and subjective difficulty.

3. The "duration of communication per unit time" defined as seconds of communication per minute correlated +0.456 with stress of time, and +0.69 with task difficulty, (both significant at p 0.01).

4. Information content per unit time, (that is, rate of information per time rather than per communication), correlated +0.63 (p 0.01) with subjective difficulty, and +0.30, (p 0.05) with stress of time.

5. If the information content of a dialogue was weighted by the time between its conclusion and the onset of the next dialogue, the resulting measure correlated +0.54 with subjective difficulty (p 0.01).

The overall conclusion would seem to be that difficulty is dependent upon the amount of spare time available; or, as the authors say:

> ". . the subjective feeling of difficulty in work processing seems to be essentially dependent on the time stress in performing the task."

Schmidt (1978) provides support for their contention that in an air traffic control SML is linked to the time spent communicating, again using a queueing theory approach.

Hacker, Plath, Richter and Zimmer (1978) report similar results from the field of process control, in a paper which unfortunately gives very few details of their methods. They recorded physiological measures such as CFF and heart rate and ratings of "mental impairment" and "emotional state". The tasks used were prose comprehension and mental arithmetic at two levels of difficulty. The harder tasks caused ratings of impairment to increase more rapidly than did the easier tasks, but there was no difference in the ratings of emotional states. Performance, physiological measures and ratings correlated +0.50. (It is rare to find high correlations between subjective, physiological and performance measures. See, for example, Hicks and Wierwille (1979) who compared five ways of measuring mental load including heart rate, primary and secondary task performance, and subjective ratings, concluding that primary task performance and

subjective scales were the methods of choice, but that task intercorrelations were on the whole small.)

Hacker et al.'s study is somewhat strange in that they appear to have asked their participants to judge the severity of the problems before attempting to solve them. They therefore must have been judging on the basis of "first impressions". As the experiment proceeded problems tended to be judged as easier than at the beginning of the experiment. The authors opt for a motivational explanation, since $\underline{d}'$ did not change. (It is not clear what they mean by this, but probably they mean the change in scores, in performance, did not alter although the subjective rating did.) They relate their results to the possesion of internal models of the types of problems, and to the skills needed to solve them, and argue that

> ". . . mental load can be reduced by transition from regulation which is achieved by the actual processing of stimuli to one based predominately on anticipation in terms of internal models carried in memory. It can be further reduced by improving those internal models."

The first steps towards a formal treatment of time stress have recently been taken by Tulga (1978). He asked operators to supervise a display which consisted of several queues in which items of different sizes and values appeared. The task was to service the queues in such a way as to maximize the net average value obtained. Tulga developed a theoretical treatment of the problem which described optimal strategies in this dynamic task in terms of rate of arrival, value, time left for task completion, etc., and found that the optimum strategy is to service items of high urgency, and only devise strategies which take account of events in the distant future when the system is relatively lightly loaded. Empirically he found that at low loads the operators appeared to treat each item separately, rather than using an overall strategy or schedule. When there was not time to complete all the tasks, and the operator was therefore overloaded, he began to develop dynamic strategies which took into account his estimate of the the likely future. Net gain increased as he used strategies, but so did subjective mental load. The imposed load was measured as:

$$L = \frac{\text{mean time to complete tasks}}{\text{meantime affordable for tasks}}$$

and at levels of L above overload subjective workload actually decreased, since the operators appear to have a less stringent criterion, and were prepared to tolerate poorer performance. That is, this was a speed-accuracy tradeoff operating in this continuous dynamic task. Daryanian and Sheridan (1980) combined values of arrival rate, available time, and duration for each task to make points on a Thurstonian scale, and found that

arrival rate of items correlated most highly with SML.

The effect of mentally loading a person is to "destabilize" the processes which use anticipation to regulate activation levels and performance. They are clearly arguing that time stress is important and that properly coded knowledge of the task can save time by increasing the efficiency of processing, and hence reduce SML. Hacker et al. claim that increased mental activity may not be accompanied by increased mental load, and cite the case of training time being reduced by 75% (sic), with less fatigue and lower ratings by the trainees most active mentally. In view of this enormous improvement it is unfortunate that no details are given.

Discussion

It seems astonishing that in the last decade there has been so little research into what makes a human feel that he is suffering from excessive mental workload. It is of course possible that there is a large data base on this subject contained in contract reports, in-house research in commercial research organizations, etc., but if so there is no trace of it in either papers or bibliographies. It appears that a number of empirically successful measures have appeared (such as the Cooper and Cooper-Harper scales) and that while these have fulfilled the requirements of assessing appropriate kinds of man-machine systems, little effort has been made explicitly to understand the origins of subjective feelings of load. This paper has expressly avoided the topic of "fatigue" because of the extreme difficulty of making sense of the reported findings, and hence the difficulty of deciding on the relevance of the work.

It seems to be that a great deal of emphasis has been placed on assessing the overall performance of man-machine systems. This, after all, is what they are designed for. If such a system fails, then a search begins for errors either in the functions of the machine or in the behavior of the human. The latter in turn are usually referred to models of information processing such as speed- accuracy tradeoff, signal detectability, information content of input or responses, decision criteria, etc., and at that point the investigation stops, with the implication that whether or not the operator feels overloaded, his feelings have little relevance to the sources of error. This may or may not be true. Certainly there are hints in the work cited above that the perceived difficulty of a task might alter the human operator's attitude toward it, and hence such things as the amount of time he would be prepared to spend, the confidence he has in his decisions, etc.

The literature on the use of the Cooper-Harper scale could have been cited at greater length, but to little purpose, since most of the investigations which use it are for the purpose of assessing the machine, not for throwing light on the human. What is really needed, and what, with the exception of the papers cited here, seems to be missing, is a systematic study of what it is that would make the CH rating given by an operator vary, other than the order of control and stability. As mentioned earlier, there is complete agreement about the latter. But supposing that they were held constant, what other variables would change the rating given by the human operator? The literature is always insistent that for rating aircraft only very experienced pilots should be used, and there are hints, which seem never to have been followed up, that individual differences might be quite considerable.

What conclusions may we draw overall about the source of the subjective feelings of being mentally loaded?

1. In manual control tasks, whether tracking tasks, vehicular or simulator control, the requirement to generate lead causes load, and the latter increases with the lead required. "Lead causes load".

2. In manual control, plant instability causes load. "Lambda causes load".

3. Below bandwidths of 1 Hz it is probable that increasing the amplitude of a forcing function increases load more than increasing bandwidth. (Note that the upper limit at which humans can perform compensatory tracking is around 1.5 Hz.)

4. In multi-loop control load is increased by heterogeneous dynamics.

5. Time pressure in the sense of the arrival of the next task demands before the last have been completed causes load.

6. The degree of precision required influences load. In particular if the number of task elements is weighted by their respective required precision, load increases with the product, (Hess, 1977).

While the above are taken for the most part from the work on manual control, it is interesting that Borg (1978b) gives a summary from the work of the Stockholm group which is very similar, allowing for the difference in the kind of tasks which they have used. According to Borg perceived difficulty is due to:

(a) the number of alternative solutions

(b) the quality of the data

(c) uncertainty about the consquences of action

(d) conflicting demands with respect to desired outcomes

(e) the need for feedback

(f) scarcity of time

(g) expenditure of energy

(h) probability of failure.

Factors such as motivation, level of aspiration, etc., should probably be added.

We obviously could relate (b) to the observation noise term in OCT models. Scarcity of time could be related to lag time constants in manual control, and to transport delay. The need for feedback and the uncertainty of consequences could be related to problems of controlling unstable plants. Scarcity of time and the conflicting demands could be related to the allocation of attention in OCT, and so on.

The overall picture, regardless of the kind of task involved, seems to be as follows. A human operator is confronted by multiple sources of information which deliver signals to him sequentially. Those signals require responses, or at least need to be processed and the results stored in memory or by means of learning. The human operator spends time extracting information, and also making decisions for action or inaction, and implementing action if required. He may be able to save time by not waiting long enough to accumulate very accurate information, or by not monitoring his responses if required accuracy is not too high. The feelings he has are related to the probability of being unable to satisfy the demands of the task in terms of accuracy, payoff, speed, etc. Some of these feelings may be due to involuntary changes in muscle tension, heart rate, etc., and some may come from perceptual or cognitive judgments that the situation is getting out of control. SML would be the experience of rising probability of failure in the near future.

It is abundantly clear that there is a real need for research in this area. For one thing, while there is a great deal of information on manual control, and some on cognitive tasks where overt responses are required, there is literally none on monitoring tasks in which humans do not exercise control but in which they monitor automatic processes to make sure that the system is within tolerable limits. A start on such work has been made by Tulga (1978) and by Yoerger (1979). The latter investigated the effect of different kinds of control on the ease of operation and efficiency of performance in a flight simulator. The lowest level of control required the operator to operate in traditional manual control mode, and the highest merely to punch in the coordinates of required way-points and altitudes. Subjective load decreased as the level of automation increased and the number of control actions decreased, the efficiency of performance increased, being linearly related to subjective workload over the range from 3 to 7 on McDonnell's (1968) scale.

From the summary above one can guess at some of the design parameters for monitoring systems which will not overload the human operator, but hard information is conspicuous by its absence in the published research literature. At present, despite its manifest importance to efficiency and safety, the only guide available is informed intuition weighted by summaries such as those above and in Johannsen et al. (1979)

References

Arnold, J.D., Johnson, R.B., & Dillow, J.D. 1973. Pitch Paper Pilot Revisited. 9th NASA Annual Conference on Manual Control. 75-81.

Ashkenas, I.L. 1966. Some Open and Closed Loop Aspects of Airplane Lateral-Direction Handling Qualities. NATO AGARDograph 533.

Ashkenas, I.L., 1972. Requirements for Airplane Normal and Failure States. Systems Technology Incorporated TR-1017-2.

Bills, A.G. 1931. Blocking: A New Principle in Mental Fatigue. American Journal of Psychology, 43, 230-245.

Borg, G. 1971. On the Problem of Perceived Difficulty. Scandanavian Journal of Psychology. 12, 249-260.

Borg, G. 1978a. Subjective Aspects of Physical Work. Ergonomics, 21, 215-220.

Borg, G. 1978b. Subjective Effort in Relation to Physical Performance and Working Capacity. In Psychology From Research to Practice, ed. Pick, H. Plenum. N.Y. Stockholm. 16.

Bratfisch, O., Borg, G., & Dornic, S. 1972. Perceived Item Difficulty in Three Tests of Intellectual Performance Capacity. Institute of Applied Psychology Stockholm, 29.

Bratfisch, O., Dornic, S., & Borg, G. 1972. Perceived Difficulty of Items in a Test of Reasoning Ability. Institute of Applied Psychology, Stockholm, 28.

Chu, Y. 1979. Adaptive Allocation of Decision-Making Responsibility Between Human and Computer in Multi-Task Situation. (in press: IEEE Transactions on Systems, Man and Cybernetics).

Cooper, G.E. 1957. Understanding and Interpreting Pilot Opinion. Aeronautical Engineering Review, 16, 47-52.

Cooper, G.E. & Harper, R.P. 1969. The Use of Pilot Ratings in the Evaluation of Aircraft Handling Characteristics. NASA TN-D-5153.

Hacker, W., Plath, H.E., Richter, P., & Zimmer, K. 1978. Internal Representation of Task Structures and Mental Load of Work: Approaches and Methods of Assesment. Ergonomics, 21, 194-197.

Hallsten, L., & Borg, G. 1975. Six Rating Scales for Perceived Difficulty. Institute of Applied Psychology Stockholm. 58.

Hess, R.A. 1973. Nonadjectival Rating Scales in Human Response Experiments. Human Factors, 15, 275-280.

Hess, R.A. 1977. Prediction of Pilot Opinion Ratings Using an Optimal Pilot Model. Human Factors, 19, 459-475.

Hess, R.A., & Teichgraber, W.M. 1974. Error Quantization Effects in Compensatory Tracking Tasks. IEEE Systems Man, & Cybernetics, SMC-4, 343-349.

Jex, H. 1979. A Proposed Set of Standardized Sub-Critical Tasks for Tracking Workload Calibration. In: Mental Workload: Theory and Measurement. ed. Moray, N. Plenum. N. Y. 179-188.

Jex, H. & Clement, W.F. 1979. Defining and Measuring Perceptual-Motor Workload in Manual Control Tasks. In: Mental Workload: Theory and Measurement. ed. Moray, N. Plenum. N.Y. 125-177.

Johannsen, G., Moray, N., Pew, R., Rasmussen, J., Sanders, A., & Wickens, C. First Steps Towards a Designer's Mental Workload Checklist. In: Mental Workload: Theory and Measurement. ed. Moray, N. Plenum. N.Y. 113-115.

Kleinman, D.L., Baron, S., & Levison, W.H. 1970. An Optimal Control Model of Human Response; Part 1.: Theory and Validation. Automatica, 6, 357-369.

Leplat, J. 1978. Factors Determining Workload. Ergonomics, 21, 143-149.

Levison, W.H. 1979. A Model for Mental Workload in Tasks Requiring Continuous Information Processing. In: Mental Workload: Theory and Measurement. ed. Moray, N. Plenum. N.Y. 189-218.

McDonnell, J.D. 1968. Pilot Rating Techniques for the Estimation and Evaluation of Handling Qualities. AFFDL-TR-76.

McDonnell, J.D. 1969. An Application of Measurement Methods to Improve the Quantitative Nature of Pilot Ratings Scales. IEEE Transactions on Man-Machine Systems, MMS-10 81-92.

McRuer, D. & Weir, D.H. 1969. Theory of Manual Vehicular Control. Ergonomics 12, 599-633.

Moray, N. (ed.) 1979. Mental Workload: Theory and Measurement. Plenum. N.Y.

Moray, N. 1979. Models and Measurement of Mental Workload. In: Mental Workload: Theory and Measurement, ed. Moray, N. Plenum, N.Y. 13-21.

Oswald, I. 1962 Sleeping and Waking. Elsevier. Amsterdam.

Parks, D. 1979. Current Workload Methods and Emerging Challenges. In: Mental Workload: Theory and Measurement. ed. Moray, N. Plenum. N.Y. 387-416.

Phillip, V., Reiche, D., & Kirchner, J. 1971. The Use of Subjective Rating. Ergonomics, 14 611-616.

Roscoe, A.H., Ellis, G.A. & Chiles, W.D. 1978. Assessing Pilot Workload. NATO AGARDograph, 233.

Rule, R.G., & Fenton, R.E., 1972. On the Effects of State Information on Driver Vehicle Performance in Car Following. IEEE Transactions on Systems, Man & Cybernetics, SMC-2, 630-637.

Schmidt, D.K., 1978. A Queuing Analysis of the Air Traffic Controllers' Workload. IEEE Transactions on Systems, Man, & Cybernetics, SMC-8, 492-498.

Sheridan, T. 1979. Personal communication.

Stave, A. 1977. The Effects of Cockpit Environment on Long Term Pilot Performance. Human Factors, 19, 503-514.

Tulga, K. 1978. Dynamic Decision Making in Multi-Task Supervisory Control: Comparison of an Optimal Algorithm to Human Behavior. Man-Machine System Laboratory Report. M.I.T.

Verplank, W.L. 1977. Is There an Optimal Workload in Manual Control? Unpublished Ph.D. Thesis, M.I.T.

Wewerinke, P.H. 1974. Human Operator Workload for Various Control Conditions. 10th NASA Annual Conference on Manual Control, 167-192.

Wewerinke, P.H. 1977. Performance and Workload Analysis of Inflight Helicopter Tasks. 13th NASA Annual Conference on Manual Control, 106-117.

Wierwille, W.W. & Williges, R.C. 1978. Survey and Analysis of Operator Workload Assessment Techniques. S-78-101. Systematics Inc., Blacksburg, Va.

Yoerger, D.R. 1979. Man-Machine Performance for a Simulated Aircraft with Multi-Level Automatic Control System. Man-Machine System Laboratory Report, M.I.T.

Yoshitake, H. 1971. Relation Between the Symptoms and the Feelings of Fatigue. Ergonomics, 14, 175-186.

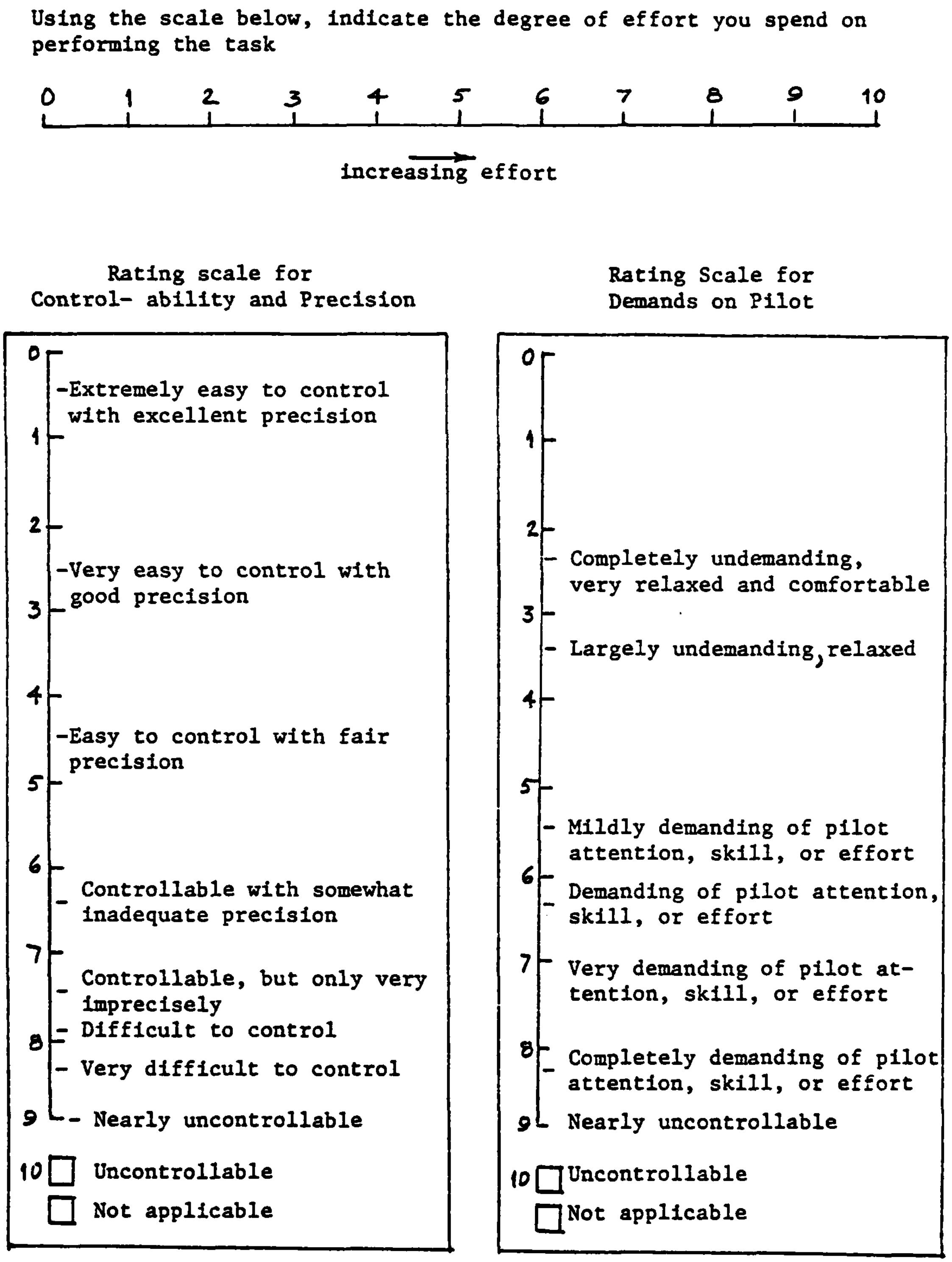


Figure 1. Three scales used by McDonnell and later Wewerinke

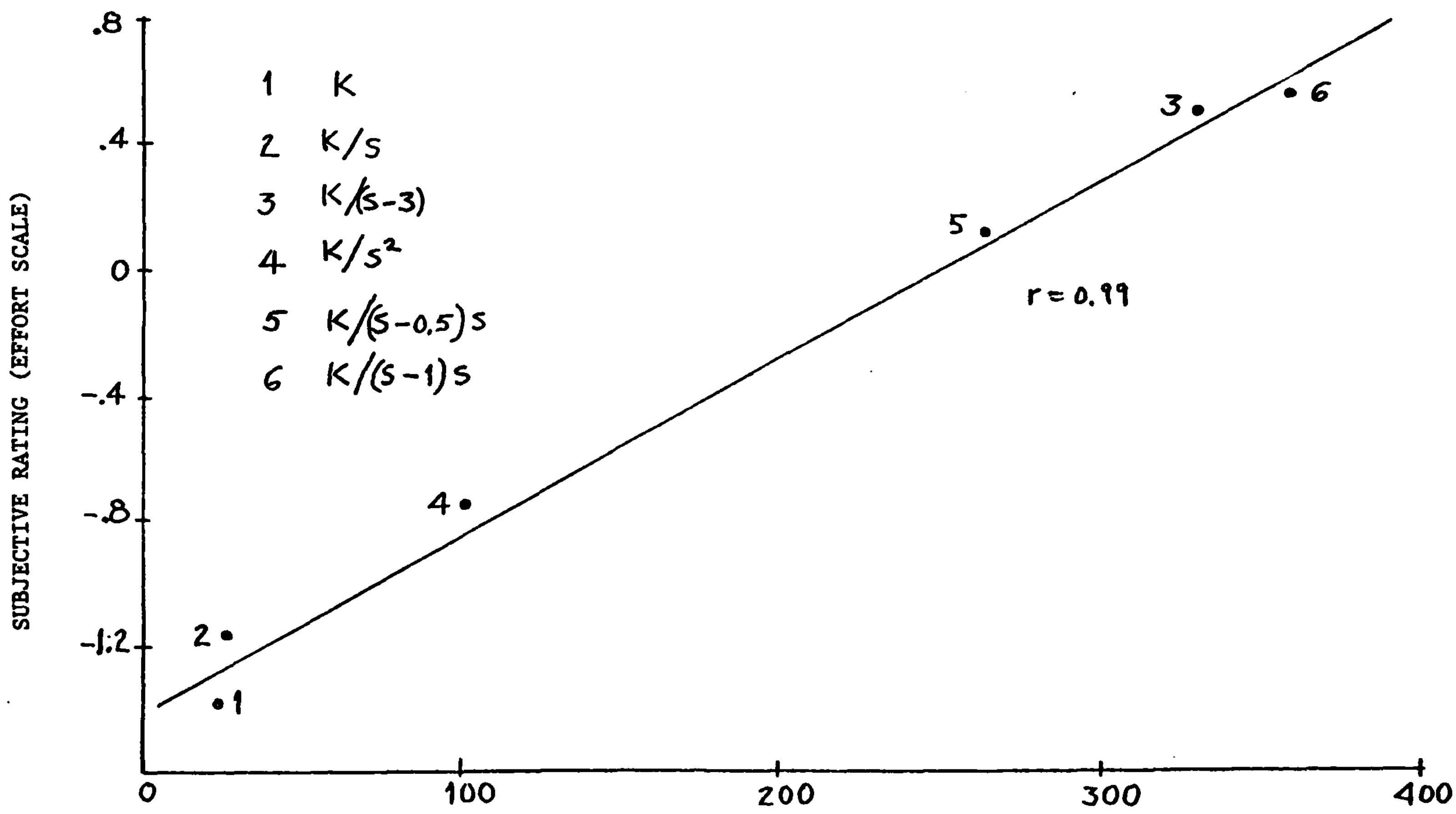
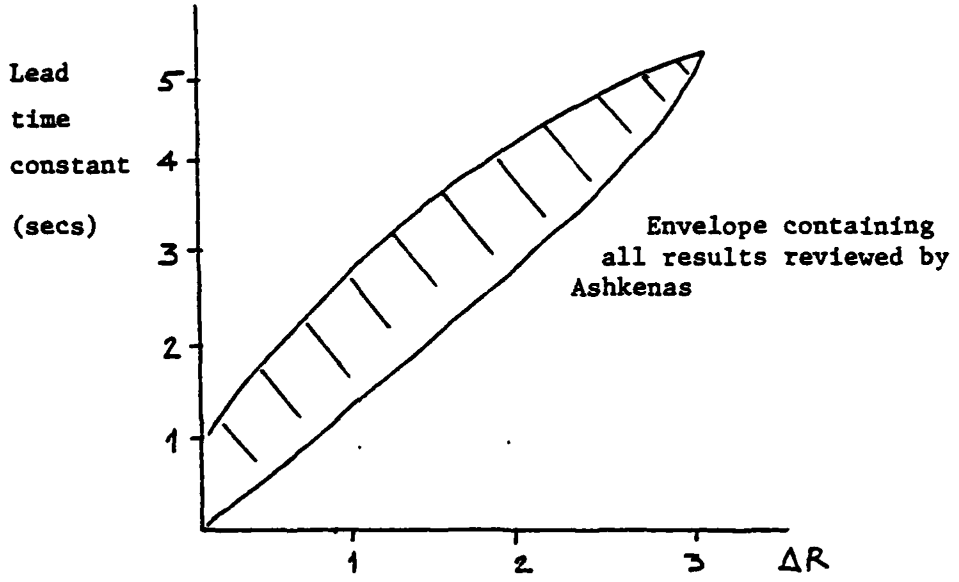
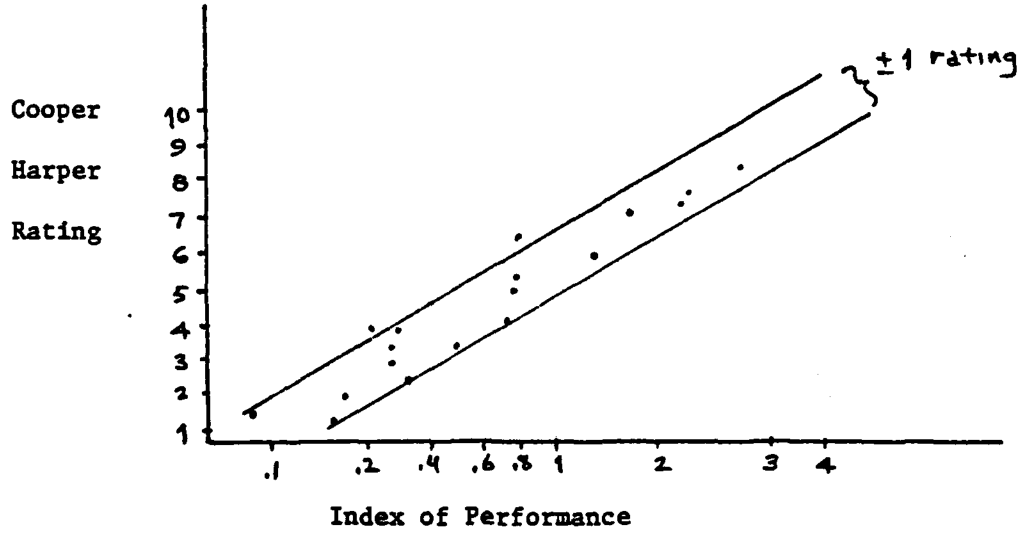


Fig. 2. The relationship between subjective ratings and computed workload, W



<u>Fig. 3</u> Decrement in Cooper rating due to lead time constant (after Ashkenas, 1966).



<u>Fig. 4</u> Cooper-Harper Ratings as a function of Index of Performance in Optimal Control Model (from Hess, 1977).

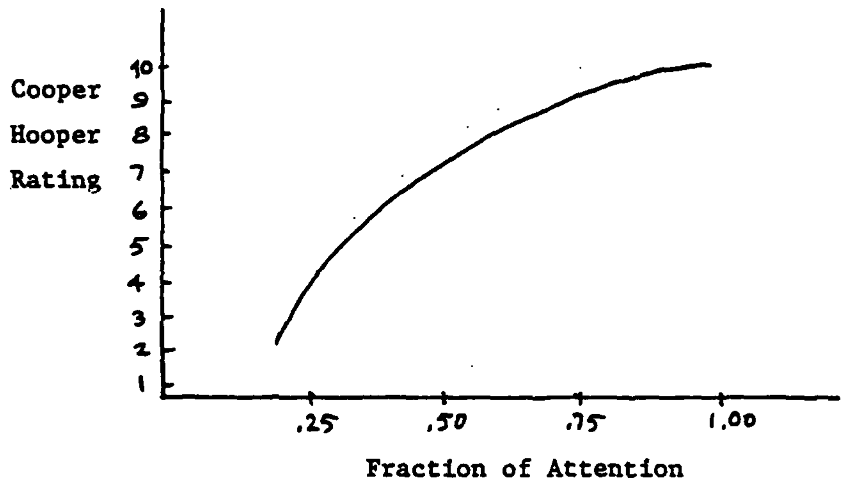


Fig. 5 Relation between Attention Allocation and Cooper-Harper Rating (from Hess, 1977)

END

DATE
FILMED

8

DTIC